## **Predicting Election Outcomes from Facial Images of Candidates Using an Unbiased Machine Learning Model**

**Raymond (Deming) Lin1\*, Jadelyn Tran2**

12Webber Academy, Calgary, AB, Canada

[raymond.lin.336@gmail.com](mailto:raymond.lin.336@gmail.com)

**Abstract**

In the past, studies on psychology have shown that humans are capable of creating instantaneous judgments of a stranger’s personality and characteristics, just based on a picture of their face. The most successful study to our knowledge has reached a 72.4% accuracy in predicting election outcomes. There have been machine learning studies that tried to replicate this success, but to our knowledge, some kind of human input and bias has always been present. We wanted to create a bias-free, independent machine learning model that only used the image of political candidates to predict their success. With no other information than a candidate’s face, we were able to achieve a 70.43% accuracy predicting election results. Not only does the different approaches in our experiment give a quantitative way to compare different types of human thinking, but it can be used as a benchmark for future research that further investigates the relationship between facial traits, human judgments, and machine learning.

*Keywords: election prediction, machine learning, classifier, transfer learning, facial images, physiognomy*

**1. Introduction (Heading 1; Use this style for heading 1)**

1.1. Background

Once every few years, the topic of election and electoral candidates floods the media. Typically, voters always try to judge candidates for who they are and how their platform would benefit them. However, no matter how they try to combat their varying level of prejudice, voters’ electoral decisions will still be heavily influenced by the candidate’s appearances (Ballew & Todorov, 2007). Studies have shown that individuals can infer various traits about a stranger just from their face (Olivola & Todorov, 2010; Olivola & Todorov, 2009; Todorov, 2018), and a lot of that subconscious interpretation of information heavily affects our electoral decision making (Ballew & Todorov, 2007). Many studies have also demonstrated that machine learning can replicate that level of human judgment. We aim to create a machine learning model with no human input and bias that can predict if a candidate wins or loses an election.

1.2. Previous Works

There have been numerous previous studies (Ballew & Todorov, 2007; Olivola & Todorov, 2010; Olivola & Todorov, 2009; Todorov, 2018) that try to test how well personality and other traits can be detected by humans by just looking at faces. In addition to lower-level traits like attractiveness (Olivola & Todorov, 2010) and emotions (Olivola & Todorov, 2009), short exposures to pictures of strangers’ faces are sufficient to predict more abstract features such as level of social class (Todorov, 2018) and election outcomes (Ballew & Todorov, 2007). For example, a seminal study of facial judgements and election outcomes was conducted by Alexander Todorov and Charles C. Ballew. Todorov and Ballew presented sixty-four test subjects with pairs of senators who were running against each other and asked them to predict who was more likely to win the election, with no time restrictions. Test subjects were only allowed to judge on the candidates that they have no familiarity with. The results indicated that the subjects successfully predicted 72.4% of the Senate races in 2006. Interestingly, when Todorov and Ballew asked participants to “deliberate and make a good judgment”, their predictive accuracy decreased; participants performed better when constrained to 250 milliseconds and thus forced to make a rapid, unreflective judgment.

These results suggest that Physiognomy, or the practice of assessing one’s character through their appearance, could have a deeper scientific reasoning that needs to be explored. Studies including Todorov and Ballew’s also suggest that the more accurate judgments are often subconscious and unintended, as the more the participants consciously tried to analyze, the worse their results were.

Inspired by the positive results of the human experiments, researchers have tried to replicate these results by building machine learning models to predict election outcomes from images of faces (Joo et al., 2015; Todorov et al., 2005; Ventura et al., n.d.). Many of these models have near-human accuracy. For example, an extensive study (Joo et al., 2015) was able to reach a 67.9% accuracy in predicting the US governor races by combining classic computer vision methods with an SVM. Specifically, the researchers trained an support vector machine (SVM) on two types of data: personality/demographic characteristics of the candidates, and physical attributes of the candidates’ portraits. To determine personality/demographic characteristics, the researchers asked human test subjects to evaluate the comparative characteristics of two candidates at a time (e.g., which candidate looks wealthier, which candidate looks older), and compiled those results into a series of perceived characteristics. To determine the physical attributes, they also split each of the training images into regions and extracted physical attributes, such as if they are smiling or if they are wearing glasses.

Despite the impressive accuracy of this method, we believe that more work is needed to fully explore this problem. For example, the researchers themselves named the categories of attributes that they want to extract, which could lead to certain physical features left ignored or under-represented. They also did not mention the logic behind determining the categories of perceived characteristics that they want to evaluate, and the logic of why those attributes were the most significant ones. Furthermore, a significant portion of the input data used for the training came from human participants to begin with; the model likely would not have achieved such a high accuracy without human contribution. In summary, previous methods such as those used in (Joo et al., 2015) relied on assumptions that introduced biases into how the algorithm made its prediction.

1.3. Our Solution

We aimed to re-examine the problem of predicting election outcomes from human faces by building a model that uses only the pixels of the facial images to lead to a result. In this way, our method can go beyond simply predicting election results and can instead be used as a benchmark for future work investigating the factors linking facial judgments to election results. More specifically, here is our approach:

1. We aim to create two approaches to machine learning models that predict a political candidate’s likelihood of winning, both using only the facial image of the candidate and no other human input and bias. In one approach, we will purely use the pixels of the images of candidates as the input to our machine learning model. With the other approach, we will simplify the image down to facial characteristics perceived by a machine learning model, and use that to predict. These two methods represent two different ways humans think when asked to predict stranger’s traits from their faces. Using those two methods as benchmarks, we can have a quantifiable measurement between human’s conscious and subconscious judgments.
2. We also try to implement two different methods of transfer learning on our machine learning model. We want to see if an encoder that is good at feature extraction with regular faces will improve the results of our machine learning model with politicians. This could help as our model may not have enough data to train with in order to reach its best performance.
3. We also want to see how much facial appearance influences election outcomes compared to other factors such as incumbency and campaign spending. We will examine the difference in accuracy with the inclusion or exclusion of the incumbency and campaign financing information in our model.

**2. Materials and Methods**

2.1. Problem Formulation

We sought to train a machine learning model to predict the outcome of an election from a facial image of a candidate. Previous studies on this subject, as far as we know, all paired politicians during their experimental or data collection process, training the model to predict which of the two candidates will win the election. This pairing was often either derived from the two candidates running against each other (Ballew & Todorov, 2007), or the two candidates with a similar perceived age (Joo et al., 2015). This logic might seem sound, as one would think that to evaluate the competence of a politician, there has to be a relative comparison. However, our experiment challenges that preconception by not using any sort of pairing in the process. All the data we used were individually selected from an array and individually evaluated, with “winner” or “loser” as the only label. This allowed us to dramatically reduce the amount of data needed to train the model due to the quadratic nature of pairwise predictions.

2.2. Data Collection

*Political Candidates Dataset*

For this experiment, we created a brand-new dataset for the facial images of politicians. We collected the names of the winning and losing politicians from the US House of Representatives elections from 2000 to 2020 (total of 9 elections) and Canadian House of Common elections from 2000 to 2021 (total of 8 elections). In total, there were around 18,000 candidates from the Canadian Parliament and 6,880 candidates from the US House of Representatives. This data formed a list of names for the politicians, and the information about what elections they won and lost also was used for labeling our dataset.

After the names of the candidates are found, we use BeautifulSoup to extract the HTML information from either google searches, specifically the knowledge panel, or from Wikipedia. We were able to find a total of 7068 facial images, one per candidate, per election.

Out of those images, we ran an automatic filter via Deepface (Serengil, 2022) to detect the face of a person in the images. This is important as later we use Deepface again to crop out parts of our image and only leave the face of the candidate. Those that did not have a detectable face in it were discarded, and 5,511 images remained. We also conducted a round of manual filtering, eliminating images that, for example, had unusual lightning or multiple faces; 4,501 images remained. For many of the politicians that were removed, we manually searched to get a different, usable photo, increasing our dataset to a total of 4,859 images. Among these images, there were candidates that lost/won multiple times, and we only kept one image in our dataset. Candidates that have both won and lost had their images eliminated entirely.

The number of “winners” and “losers” has to be the same while training the program. During training, 2,018 unique images of candidates (1 image per candidate) were actually used, 1,009 for each category, with some “winner” images left unused (randomly selected during each train/test split). Using Deepface, these images were converted to grayscale and cropped to 48 by 48 pixels centered in their face, without any information of their background or clothing.

*Incumbency and Campaign Financing Data*

Although the primary focus of our study is on facial images, we wanted to see how much incumbency and campaign financing information influences our predictions. We used BeautifulSoup to collect data from Ballotpedia, and using the information found by web scraping the HTML code, we were able to gather the information of 609 US candidates. These candidates that have this information were trained again (later fully explained under models - Method 3), this time with incumbency and financial spending as additional input variables.

*Examples*

Figure 1a & 1b. This is an example of a candidate in the “loser” category (Val Demings, Florida 11th District). These are the original image we collected and the cropped image used for training, respectively.

##### 

Figure 2a & 2b. This is an example of a candidate in the “winner” category (Pete Stark, California 13th District). These are the original image we collected and the cropped image used for training, respectively.

2.3. Models

To fully understand our problem, we created four methods in total. We manipulated the specific input variables and the training process in order to have a more comprehensive suite of results. This furthers our understanding of this topic, and gives us a more holistic perspective. For our methods, we used a classifier for images, classifier for images with an autoencoder, logistic regression with Deepface.analyze (Serengil, 2022), and classifier with Deepface.represent. We developed all of our coding with Python and Google Colab (“Google”, 2017). Specifically, we used PyTorch (“PyTorch”, n.d.) for all of the programming with the machine learning models. We also used Deepface (Serengil, 2022), a python library that has extensive functions for facial recognition and analysis.

*Method 1 – Pixel-Based Classifier*

Using PyTorch (“PyTorch”, n.d.), we developed a neural network with linear layers that classified our given inputs into either the “winner” or the “loser” category. This classifier was trained only using normalized arrays of the 48x48 pixel, gray scale images, with everything but the politician’s face cropped out of the photo. The model was trained with BCELoss as the loss function and a learning rate of 0.0001, until it reached a point of diminishing returns (around 750-1000 epochs).

In the classifier, there are three linear layers in total. The first layer accepts 2,304 inputs (48 by 48 pixels), and has 256 outputs. The second layer decreases that to 128 outputs, and the third layer decreases it to 1 (a binary unit indicating if the candidate is a winner or a loser). We used ReLU as the activation function of the first two layers, and Sigmoid for the last layer.

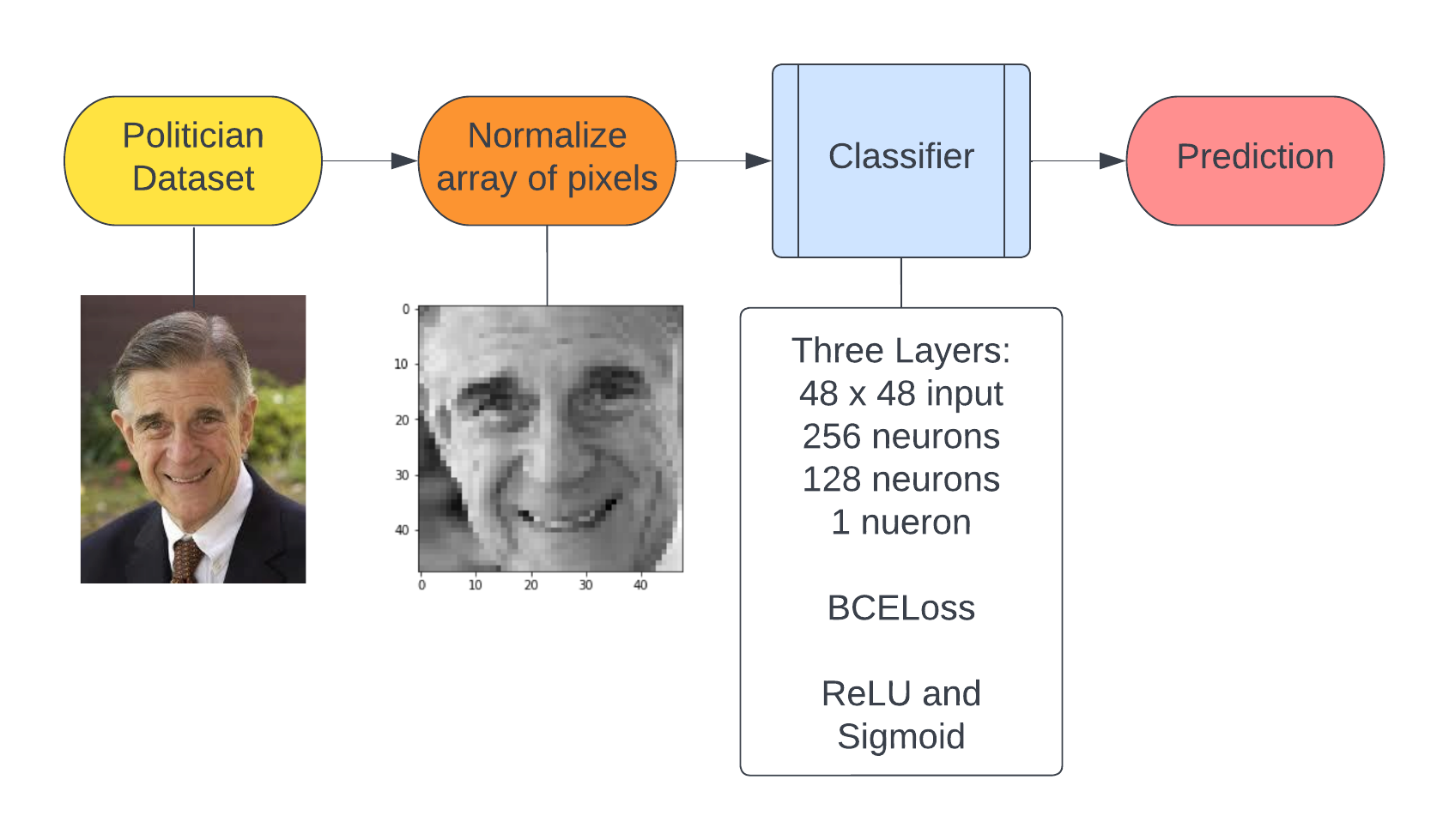


Figure 3. This is the general structure of Method 1. The classifier uses pixels of facial images to determine a result.

*Method 2a – Pixel-Based Classifier with Autoencoder*

Due to the limited size of the training dataset, we next sought to improve the performance of the model with pre-training. We applied transfer learning to the classifier with the use of an autoencoder, creating an autoencoder and training it on an independent dataset (Arora, 2020) of regular faces, without any labels. We then used the information contained in the latent space of the autoencoder and reapplied that to our classifier in Method 1 to see if it improved the accuracy of our training.

The autoencoder was trained to extract features and compress that information as much as possible, while still being able to recreate the image with that compressed information. Our autoencoder is a Convolutional Neural Network that has three Conv2d layers for encoding, and three ConvTranpose2d layers for decoding. Our activation function was three ReLU functions for encoding, and two ReLU and one Sigmoid function for decoding. We saved the weights of the autoencoder, and applied them to our classifier.

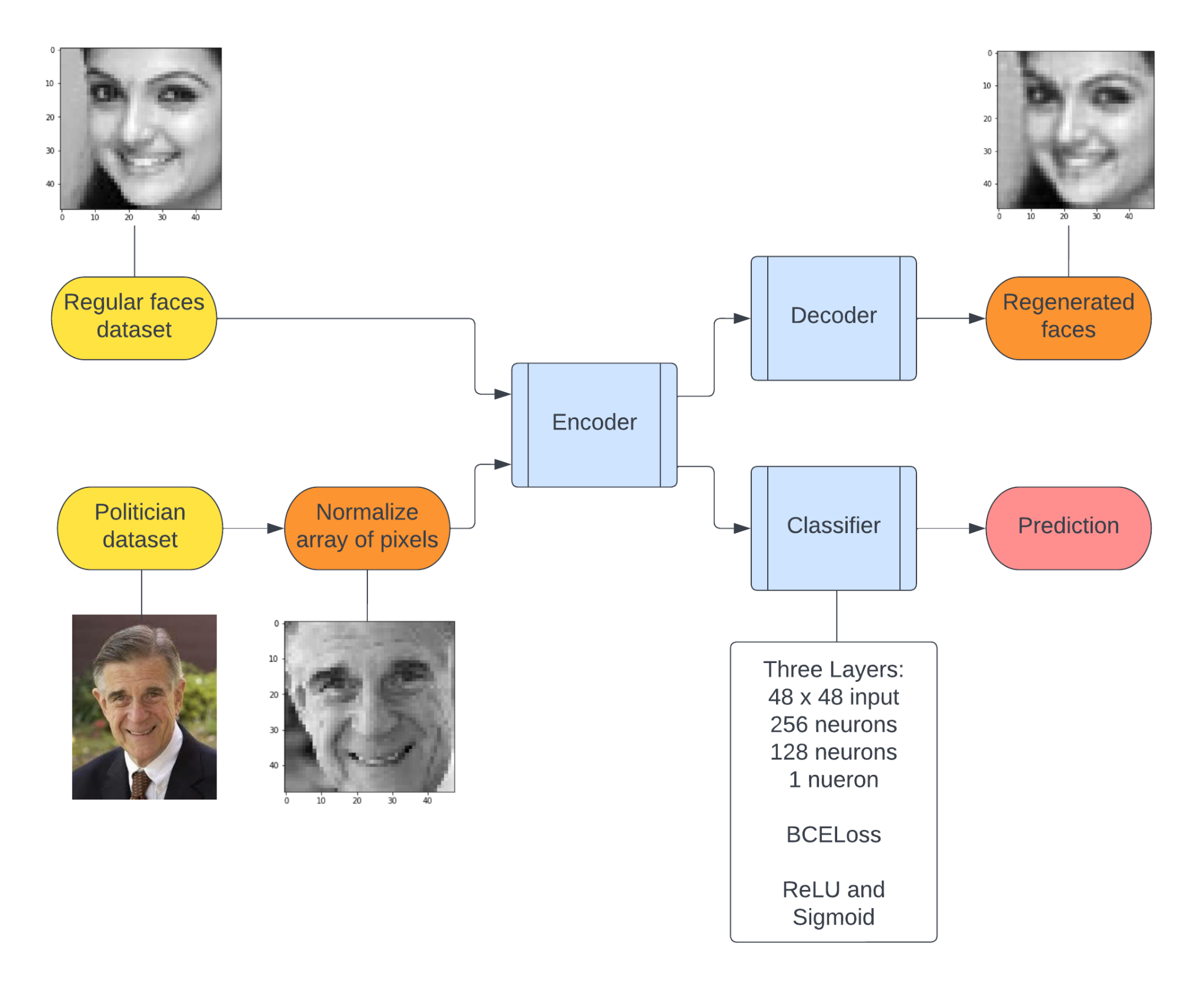


Figure 4. This is the general structure of Method 2a. The difference between Method 1 and 2a is that Method 2a uses an autoencoder to implement transfer learning, and its encoder is used with the classifier to improve its feature extraction abilities

*Method 2b – Pixel-Based Classifier with Deepface Representations*

To implement another method of transfer learning, we used a pre-trained model to extract features from the candidates’ images. We used the returned values of Deepface.represent (Serengil, 2022) as the input variable for this method. Deepface.represent is meant to be used for facial recognition by representing faces with vector embeddings. We used these embeddings as the input for a classifier. This classifier is similar to the classifier used in Method 1 and 2a; the difference is that since the input variables were changed, we had changed the expected input of the first layer as well.

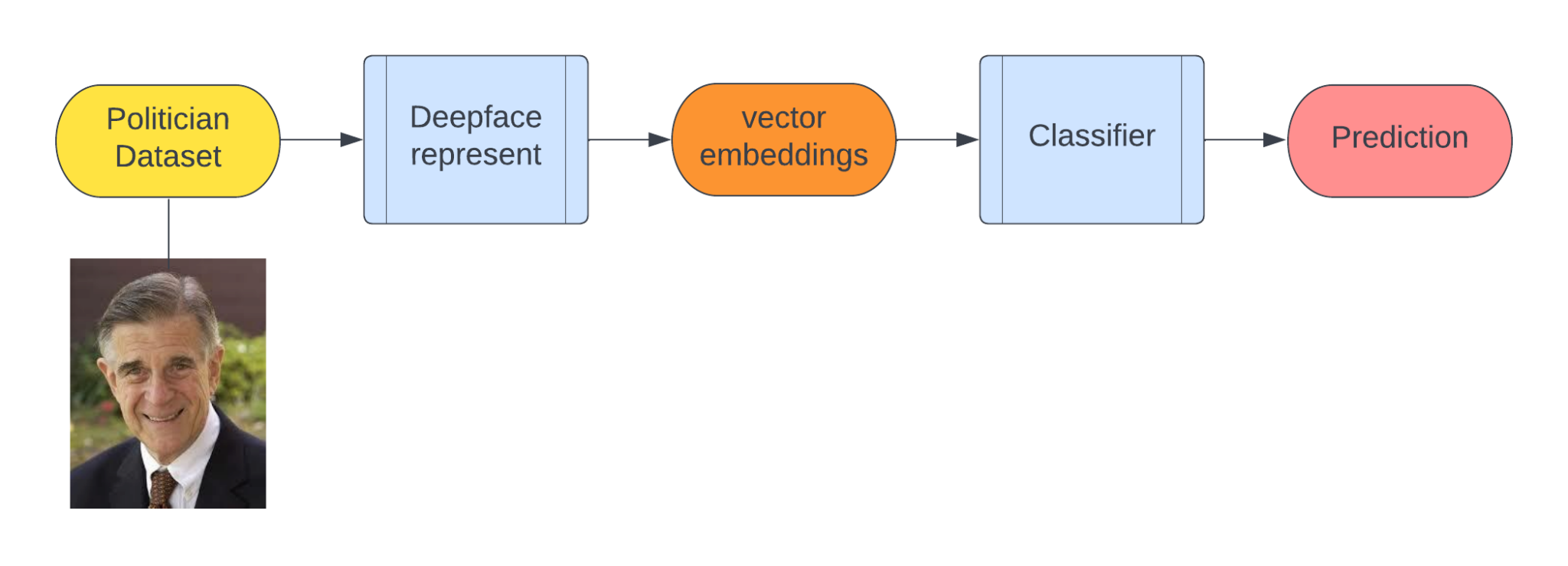


Figure 5. This is the general structure of Method 2b. Just like Method 2a, this is another implementation of transfer learning using Deepface.represent

*Method 3 – Demographic-Informed Classifier*

As a comparison to the prior 2 methods, which predict election outcomes from facial images directly, for this method we trained a model using demographic/facial expression information. We used the Python library Deepface, and specifically the Deepface.analyze function, to generate these features. We fed the function the images of the politicians, and it returned a series of perceived values of the politicians, such as their age, composition of race, likelihood of certain emotions, gender, etc. These values were used as input variables for a Logistical Regression model from SkLearn (“Learn: Machine”, n.d.). Later on, we also retrained this with the incumbency status and the amount spent in campaign financing as additional input variables.

Deepface contains pre-trained frameworks that can analyze attributes of facial images. Deepface.analyze is a function that can return multiple attributes from a facial image. We are using all the information that it returns, which includes likelihood of certain emotions (angry, neutral, happy, etc.), likelihood of race (White, Latino, Asian…), age and gender.

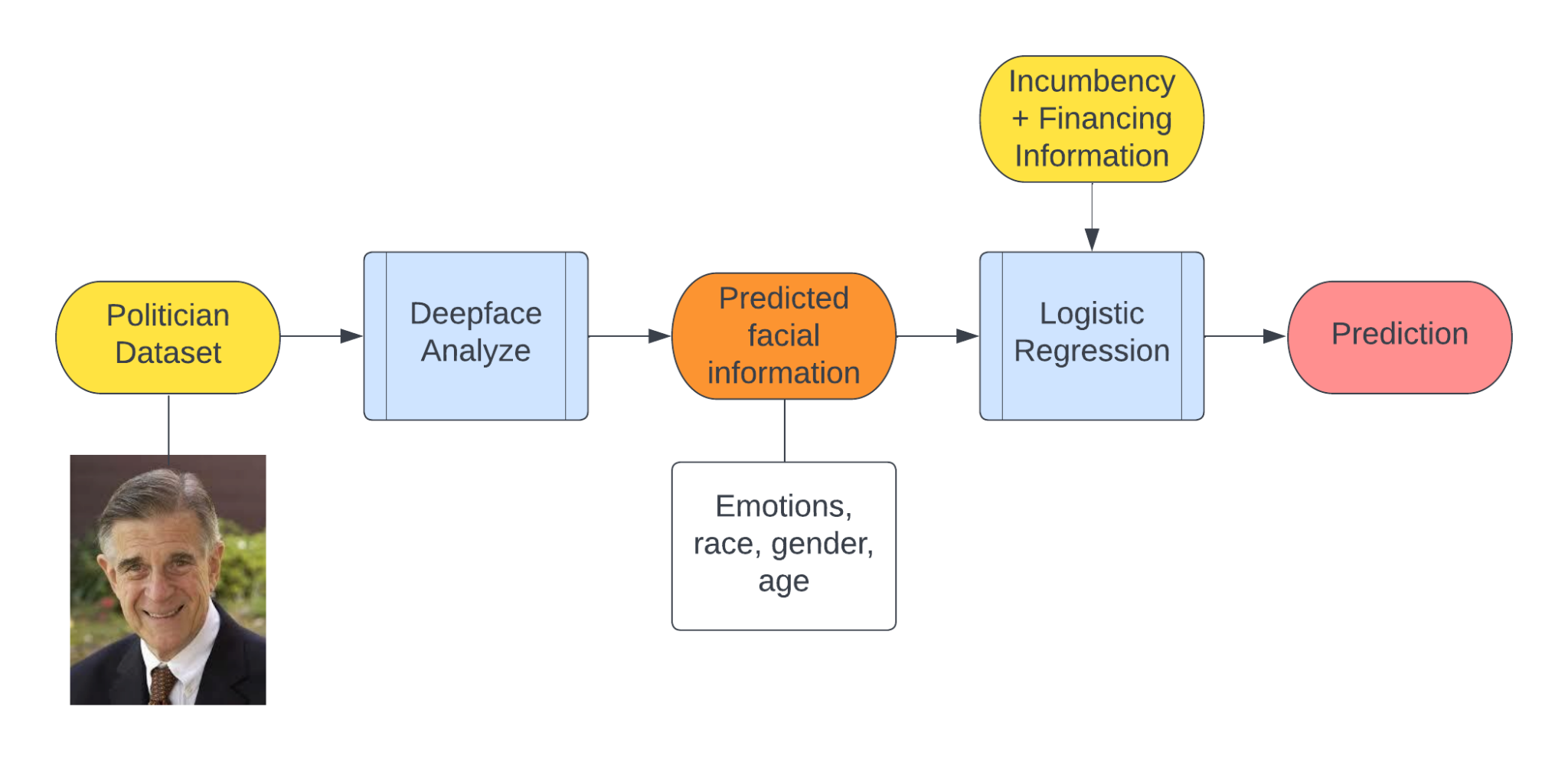
**

Figure 6: general structure of Method 3. The perceived traits from Deepface.analyze are used as input for a logistic regression model that predicts a result

**3. Results**

We evaluated the testing dataset’s accuracy and the ROC Curve. The following is our result.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Method 1 | Method 2a | Method 2b | Method 3 |
| Training accuracy (%) | 94.20 | 95.36 | 61.16 | 60.45 |
| Testing Accuracy (%) | 70.43 | 66.96 | 16.57 | 59.95 |
| AUC | 0.77 | 0.74 | 0.37 | 0.61 |

3.1. Results for Method 1, 2a, 2b

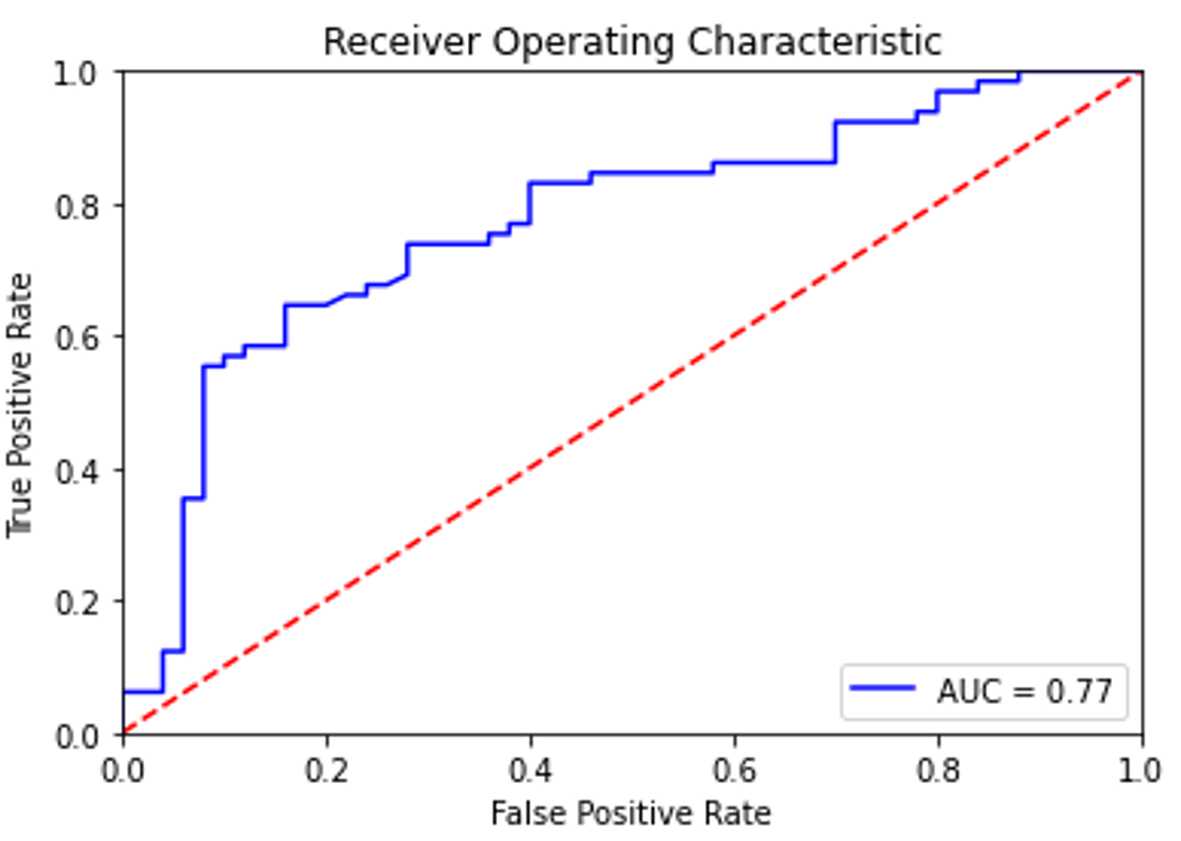


Figure 7: ROC curves for Method 1

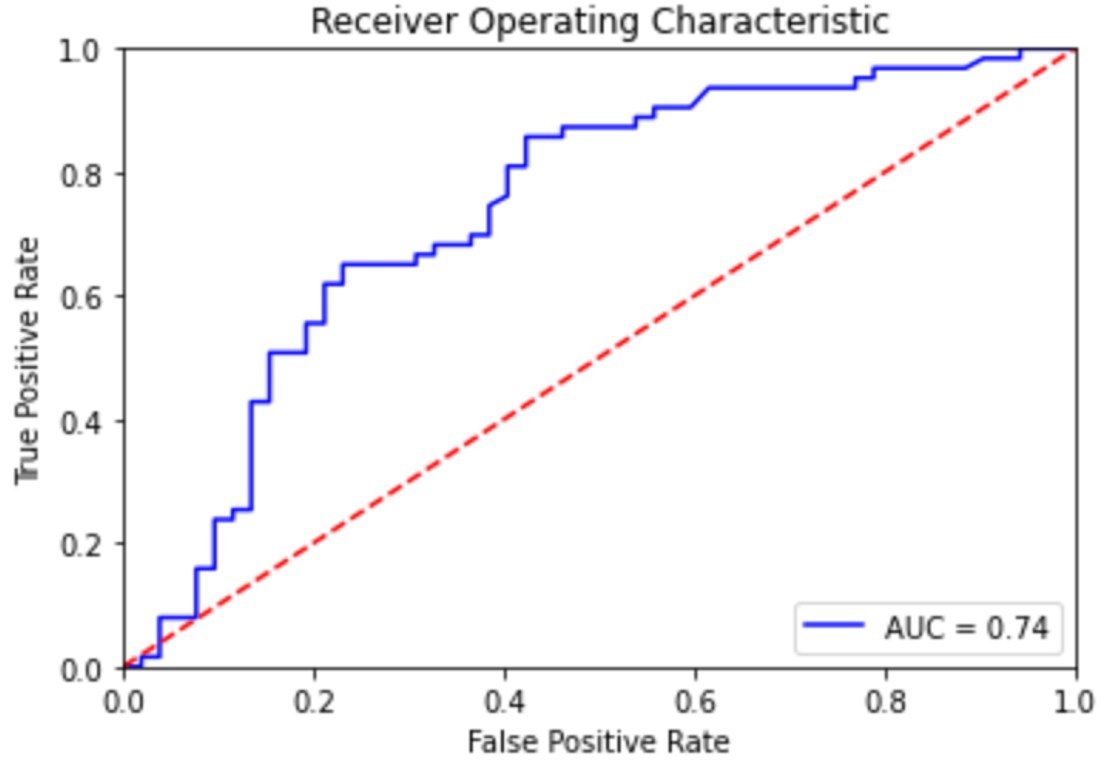
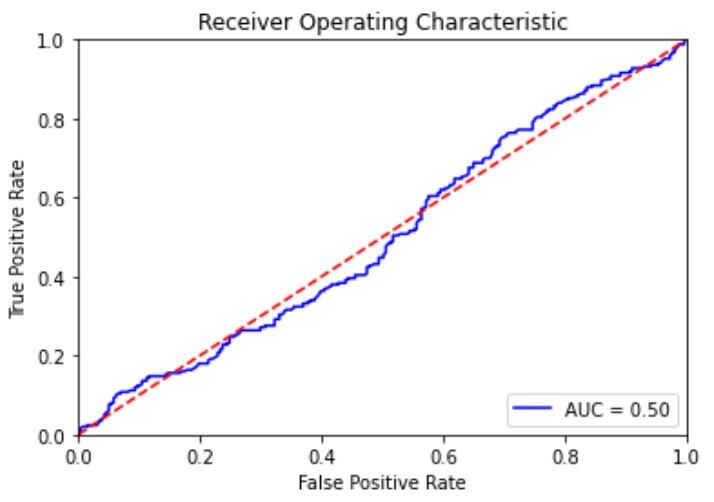
****

Figure 8a & 8b: ROC curves for Method 2a and Method 2b respectively

Confusion Matrix for Method 1

|  |  |  |
| --- | --- | --- |
|  | Actually Winners | Actually Losers |
| Predicted Winners | 44 | 6 |
| Predicted Losers | 28 | 37 |

Confusion Matrix for Method 2a

|  |  |  |
| --- | --- | --- |
|  | Actually Winners | Actually Losers |
| Predicted Winners | 36 | 16 |
| Predicted Losers | 22 | 41 |

From these results, it is clear that Method 1, with its 70.43% accuracy is the best approach. Method 2a is extremely close, with a 3.47% worse testing accuracy. It correctly predicted 4 candidates less than Method 1. However, it is interesting how Method 2a provided a better accuracy for predicting Losers. Method 2b, however, is very unsuccessful. It only achieved a 16% testing accuracy, which is worse than simply guessing with a 50% chance. Its AUC value is unimpressive as well.

3.2. Results for Method 3

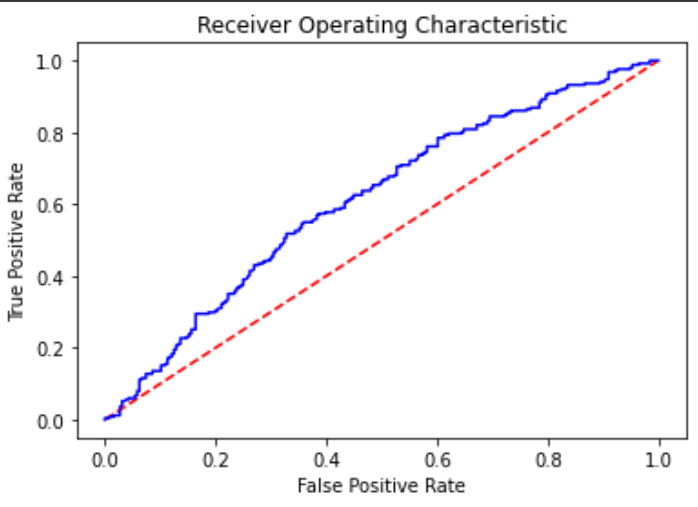
****

Figure 9: ROC curves for Method 3

Method 3 is relatively successful, achieving a 59.95% testing accuracy, with an AUC value of 0.61. It is not close to the success with Method 1, but its accuracy is significantly higher than 50%, the accuracy of making predictions with random guesses.

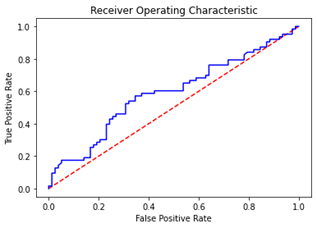
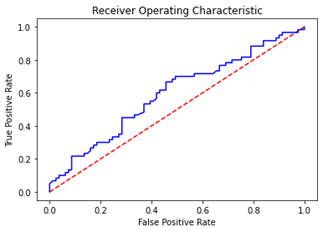
 

Figure 10A & 10B: ROC curves for variations of Method 3. These are results with the inclusion / exclusion of incumbency information. (330 politicians)

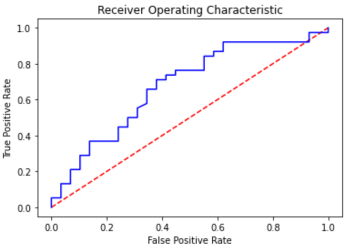
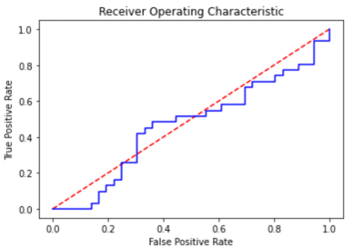
 

Figure 10C & 10D: ROC curves for variations of Method 3. These are results with the inclusion / exclusion of campaign financing information. (133 politicians)

As stated before, we retested Method 3 by adding incumbency and amount spent on the campaign as additional parameters for the input of our linear regression model.

The incumbency and financing information was only available for some candidates (330 for incumbency and 133 for financing). Figure 10B and 10D are baseline testing that have no information besides a candidate’s face, while figure 10A and 10C are the results after incumbency / financial information is added to each group. Both with incumbency and financing, it shows that the addition of this information greatly helps with the results in the respective groups.

The training with financing information as the only additional information besides the ones provided by Deepface.analyze has the best ROC curve, with the highest AUC value out of the four (0.677). The training with incumbency information is close behind in those terms, but has a higher accuracy (highest of the four) of 58.79%.

**4. Discussion**

Our model from Method 1, while only using the pixels of the images of a candidate’s face, was able to reach 70.43% accuracy. It had no information on the political affiliation of the candidate, and no information of the political inclination of the region they are running in. Our model was able to accomplish this without any other information, completely disregarding the competency, incumbency, or campaign spending of the candidate. In contrast to prior work (Joo et al., 2015), our model had no human input on the dataset as it was trained solely with a “winner/loser” label.

These results are quite significant, as it suggests that Artificial Intelligence on its own, with no human influence, was able to predict election outcomes from faces. Since our model had no other information to rely on, it demonstrates how the facial appearance of a candidate is a heavy contributing factor in the minds of a voter.

To make up for the relatively small size of our dataset, we also attempted two methods of transfer learning (Method 2a & 2b). We used generic data of faces to help with feature extraction, which we then can apply onto our election prediction model. Neither approach was more successful than Method 1 & Method 3. This suggests that the factors that are important for judging faces for elections are somewhat unique, and not the exact same as facial information that might be reflected in our process of transfer learning. A conclusion that we might reach from this is that on a dataset specifically targeted for a purpose, a large amount of data may not be necessary to achieve a significant result.

In previous human psychological studies on this topic, test subjects often only have a second or less to make a prediction. For most humans, that is not enough time to consciously analyze the candidate's faces and infer traits. However, studies have shown that the shorter the time taken, the better the prediction (Ballew & Todorov, 2007). The explanation given was that humans’ subconscious judgments or their “intuitions” are incredibly good at making decisions. In these cases, humans’ subconscious made even better predictions than conscious analysis.

There is still debate about exactly how humans are able to do this, and the true nature of these seemingly amazing judgments are yet to be fully understood. Are these judgments values that can be easily represented by numbers and categories of perceived emotions or traits, or are they reflections of nebulous, underlying factors that cannot be simply understood with a few words? Our results provide a quantifiable way to compare these two possibilities.

Method 1 acts as a representation of human judgments that include underlying factors, while Method 3 only uses quantifiable and simplified traits to form a prediction. Method 3 only reached 59.41% accuracy. However, it shows that, to a certain extent, simple and quantifiable traits such as emotions, race and age can still be incredibly useful information to analyze political candidate’s success. However, since Method 1 achieving the highest accuracy of 70.43% shows that there are factors unaccounted for in Method 3, and some useful traits cannot be successfully simplified or represented with understandable and basic categories.

In this project, we used Method 1 and Method 3 as representations of two ways of human thinking. Method 1 (by using all the pixels of a facial image) represents intuition: the subconscious, unreflective judgments that cannot be represented with a few words or values. Method 3 (by simplifying facial information into emotions, age, gender) represents consciousness: when humans intentionally try to analyze information and categorize into specific and familiar groups. Using machine learning, we have provided a new way to make a quantifiable assessment between the two methods of human judgment. With further integration into existing systems of political forecasting, our model’s ability to assess voter bias could be crucial to improving election predictions. Furthermore, our work is useful for any researchers that want to build upon our findings, use it as a benchmark, and further explore the nuances of human judgments and machine learning. Future studies based on different theories and principles can also use our model to compare with to determine which features are truly important for predicting election outcomes.

**Acknowledgment**

Majors thanks to Tyler Giallanza for his guidance and mentorship.

**References**

Arora, N. (2021). Age, Gender and Ethnicity (Face Data) csv. Kaggle.

Ballew, C. C., & Todorov, A. (2007, November 13). *Tipping elements in the Earth's climate system | PNAS*. Predicting political elections from rapid and unreflective face judgments. Retrieved January 2, 2023, from <https://www.pnas.org/doi/10.1073/pnas.0705414105>

Google Colaboratory. (2017). computer software. Retrieved January 2, 2023.

Joo, J., Steen, F. F., & Zhu, S.-C. (2015, December). Automated Facial Trait Judgment and Election Outcome Prediction: Social Dimensions of Face. International Conference on Computer Vision.

Olivola, C. Y., & Todorov, A. (2010, January 23). Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting.

Olivola, C., & Todorov, A. (2009, May 5). The Look of a Winner. Retrieved January 2, 2023, from <https://www.scientificamerican.com/article/the-look-of-a-winner/>

PyTorch. (n.d.). Version (1.13.1). *From Research To Production*. Retrieved January 2, 2023, from <https://pytorch.org/>.

scikit-learn. (n.d.). Version (1.2.0). *scikit-learn Machine Learning in Python*. Retrieved January 2, 2023, from <https://scikit-learn.org/stable/>.

Serengil, S. I. (2022, May 10). DeepfaceVersion (0.0.75). *deepface 0.0.75*. PyPI. Retrieved January 2, 2023, from <https://pypi.org/project/deepface/>.

Todorov, A. (2018, May 14). *Can we read a person's character from facial images?* Scientific American Blog Network. Retrieved January 2, 2023, from <https://blogs.scientificamerican.com/observations/can-we-read-a-persons-character-from-facial-images/>

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005, June 10). Inferences of Competence from Faces Predict Election Outcomes. Washington D.C.; American Association for the Advancement of Science.

Ventura, C., Masip, D., & Lapedriza, A. (n.d.). Interpreting CNN Models for Apparent Personality Trait Regression. Institute of Electrical and Electronics Engineers.